REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	June 2001	Annual Summary (1 Jun 00 - 31 May 01)

4. TITLE AND SUBTITLE

Computer-aided Diagnosis of Digital Mammograms

5. FUNDING NUMBERS

DAMD17-00-1-0197

6. AUTHOR(S)

Yulei Jiang, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

The University of Chicago
Chicago, Illinois 60637
E-Mail: y-jiang@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

20011130 038

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

14. SUBJECT TERMS

breast cancer

15. NUMBER OF PAGES

32

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

AD_____

Award Number: DAMD17-00-1-0197

TITLE: Computer-aided Diagnosis of Digital Mammograms

PRINCIPAL INVESTIGATOR: Yulei Jiang, Ph.D.

CONTRACTING ORGANIZATION: The University of Chicago
Chicago, Illinois 60637

REPORT DATE: June 2001

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

TABLE OF CONTENTS

INTRODUCTION

The long-term goal of our research is to develop computer-aided diagnosis (CAD) techniques for improving the detection and diagnosis of breast cancer. The hypothesis to be tested in the present project is that radiologists' ability to differentiate malignant from benign breast lesions can be improved by integrating radiologists' perceptual expertise in the interpretation of mammograms with the advantages of automated computer classification. This project has 3 objectives:

1. To combine radiologist-extracted Breast Imaging Reporting and Data System (BI-RADS) features with image features extracted by a computer to classify malignant and benign clustered microcalcifications in mammograms.
2. To optimally combine radiologists' diagnosis with the result of computer classification.
3. To optimize computer classification for full-field digital mammograms.

BODY

1. Analysis of clinical benefits of CAD

We analyzed data obtained in a previous observer study to find potential clinical benefits from CAD in addition to what has already been demonstrated [1]. In this observer study, 10 radiologists reviewed mammograms of 104 patients both without and with a computer aid that was designed to help them differentiate malignant from benign clustered microcalcifications in mammograms. Previously, we demonstrated that the computer aid helped the radiologists to improve diagnostic accuracy as measured by A_z (area under a receiver operating characteristic [ROC] curve) [1]. Specifically, the computer aid helped each of the radiologists, on average, to recommend 14% more biopsies for malignant lesions and to recommend 10% fewer biopsies for benign lesions. In the present analysis, we demonstrated that the computer aid helped reduce substantially the variabilities in radiologists' interpretation of the mammograms. In addition, the computer aid helped radiologists to improve diagnostic accuracy to a greater extent compared to independent double readings, i.e., the interpretation of the same mammograms by two different radiologists independently and the subsequent combination of their diagnoses. Parts of this work were presented at an annual meeting of the American Association

of Physicists in Medicine (AAPM) [2] and a Scientific Assembly and Annual Meeting of Radiological Society of North America (RSNA) [3]. In addition, a publication has resulted in collaboration with Dr. Robert Wagner who performed a separate theory-based analysis of our observer-study data to demonstrate the reduction in variability due to CAD in the interpretation of mammograms (see reprint in the Appendix) [4].

2. Comparison of BI-RADS lesion descriptors and computer-extracted image features

A study has been ongoing to compare computer classification of breast lesions as malignant or benign based on BI-RADS lesion descriptors [5, 6] and based on computer-extracted image features that we have described previously [7, 8]. Our goal in this study was to identify the relative strengths and weaknesses of the two different computer classification methods and to improve computer classification by developing new computer image feature-extraction techniques that correspond to, and that can be used to substitute for, important BI-RADS lesion descriptors. In this study, we included both clustered microcalcifications and masses, even though we originally proposed to study only clustered microcalcifications. We have collected a total of 209 cases for this study, 123 of which contain clustered microcalcifications and 86 contain masses. There are 85 malignant lesions and 124 benign lesions in this database. All cases include original mammograms in the standard and magnification views. Currently, we are collecting data on BI-RADS lesion descriptors in an observer study.

3. "Optimal" combination of radiologists' and a computer's diagnostic assessment

We have developed a method for combining quantitative diagnostic assessments made by radiologists and those made by a computer aid. This method was based on a bivariate binormal model that was also used in ROC analysis. This method takes into account the individual accuracy of the radiologist and the computer aid, and the correlation between their diagnostic assessments. We applied this method to data obtained previously from an observer study and found that the results obtained using this method was better than the results that radiologists achieved by using the computer aid in an *ad hoc* way. The average A_z value increased from 0.75 to 0.79. The improved A_z value was close to the performance of the computer alone ($A_z = 0.80$). This work was presented at the SPIE [9]. A conference proceeding is included in the Appendix [10].

4. CAD in small-field digital mammograms

We conducted a study of computer classification of malignant and benign clustered microcalcifications in small-field digital mammograms. Our goal was to develop CAD for full-field digital mammograms based on our techniques developed previously on digitized screen-film mammograms. The purpose of this work was to evaluate the feasibility of applying our existing computer technique to mammograms acquired with a digital detector without extensive modifications. We analyzed 79 lesions that were biopsied. Of these, 33 lesions were malignant and 46 were benign. Because each case normally consisted of more than one image, we analyzed a total of 176 images, of which 56 were of the malignant lesions and 120 were of the benign lesions. The computer analysis achieved an A_z value of 0.84 for the 176 images and 0.90 for the 79 lesions. In comparison, radiologists who evaluated these lesions prior to the biopsies achieved an A_z value of 0.76 for the 79 lesions. This study demonstrated the potential of our computer technique to classify accurately clustered microcalcifications in mammograms acquired with a digital detector as malignant or benign. This study was presented at the 5th International Workshop on Digital Mammography [11]. A conference proceeding is included in the Appendix [12].

KEY RESEARCH ACCOMPLISHMENTS

- Analysis of potential clinical benefits of CAD of malignant and benign breast lesions.
- Work-in-progress on a comparison of BI-RADS lesion descriptors provided by radiologists and computer-extracted image features for computer classification of breast lesions as malignant or benign.
- Development of a novel method for the "optimal" combination of quantitative diagnostic assessments made by a radiologist and that made by a computer.
- Investigation of computer classification of malignant and benign clustered microcalcifications in small-field digital mammograms.

REPORTABLE OUTCOMES

1. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Multiple benefits of computer-aided diagnosis (CAD) in the diagnosis of malignant and benign breast lesions. Presented at World Congress on Medical Physics and Biomedical Engineering, July, 2000.
2. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Three potential benefits of computer-aided diagnosis (CAD) in breast cancer diagnosis. Chicago, Illinois: the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, 2000.
3. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 8:605-615, 2001.
4. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. Presented at SPIE's International Symposium: Medical Imaging 2001, February, 2001.
5. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. Proc. SPIE 4324:177-183, 2001.
6. Jiang Y, Nishikawa RM, Venta LL, Maloney MM, Giger ML. Computer classification of malignant and benign microcalcifications in small-field digital mammograms. Presented at 5th International Workshop on Digital Mammography, June, 2000.
7. Jiang Y, Nishikawa RM, Venta LL, Maloney MM, Giger ML. Computer classification of malignant and benign microcalcifications in small-field digital mammograms. In: IWDM 2000 5th International Workshop on Digital Mammography (Yaffe MJ eds.). Medison, WI: Medical Physics Publishing, pp. 237-242, 2000.

CONCLUSIONS

We have made progress toward all 3 objectives of this project. The research results are positive and support the continuation of this project.

REFERENCES

1. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 6:22-33, 1999.
2. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Multiple benefits of computer-aided diagnosis (CAD) in the diagnosis of malignant and benign breast lesions. Presented at World Congress on Medical Physics and Biomedical Engineering, July, 2000.
3. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Doi K. Three potential benefits of computer-aided diagnosis (CAD) in breast cancer diagnosis. Chicago, Illinois: the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, 2000.
4. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: the case of unequal variance structures across modalities. Acad Radiol 8:605-615, 2001.
5. American College of Radiology (ACR). Breast imaging reporting and data system (BI-RADSTM). Vol. Third Edition ed. Reston, VA: American College of Radiology, pp. 1998.
6. Baker JA, Kornguth PJ, Lo JY, Floyd CEJ. Artificial neural network: improving the quality of breast biopsy recommendations. Radiology 198:131-135, 1996.
7. Jiang Y, Nishikawa RM, Wolverton DE, Metz CE, Giger ML, Schmidt RA, Vyborny CJ, Doi K. Malignant and benign clustered microcalcifications: automated feature analysis and classification. Radiology 198:671-678, 1996.
8. Huo Z, Giger ML, Vyborny CJ, Wolverton DE, Schmidt RA, Doi K. Automated computerized classification of malignant and benign masses on digitized mammograms. Acad Radiol 5:155-168, 1998.
9. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. Presented at SPIE's International Symposium: Medical Imaging 2001, February, 2001.
10. Jiang Y, Metz CE. An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis. Proc. SPIE 4324:177-183, 2001.
11. Jiang Y, Nishikawa RM, Venta LL, Maloney MM, Giger ML. Computer classification of malignant and benign microcalcifications in small-field digital mammograms. Presented at 5th International Workshop on Digital Mammography, June, 2000.
12. Jiang Y, Nishikawa RM, Venta LL, Maloney MM, Giger ML. Computer classification of malignant and benign microcalcifications in small-field digital mammograms. In: IWDM 2000 5th International Workshop on Digital Mammography (Yaffe MJ eds.). Medison, WI: Medical Physics Publishing, pp. 237-242, 2000.

Components-of-Variance Models for Random-Effects ROC Analysis:

The Case of Unequal Variance Structures Across Modalities[1]

Sergey V. Beiden, PhD, Robert F. Wagner, PhD, Gregory Campbell, PhD, Charles E. Metz, PhD, Yulei Jiang, PhD

Rationale and Objectives. Several of the authors have previously published an analysis of multiple sources of uncertainty in the receiver operating characteristic (ROC) assessment and comparison of diagnostic modalities. The analysis assumed that the components of variance were the same for the modalities under comparison. The purpose of the present work is to obtain a generalization that does not require that assumption.

Materials and Methods. The generalization is achieved by splitting three of the six components of variance in the previous model into modality-dependent contributions. Two distinct formulations of this approach can be obtained from alternative choices of the three components to be split; however, a one-to-one relationship exists between the magnitudes of the components estimated from these two formulations.

Results. The method is applied to a study of multiple readers, with and without the aid of a computer-assist modality, performing the task of discriminating between benign and malignant clusters of microcalcifications. Analysis according to the first method of splitting shows large decreases in the reader and reader-by-case components of variance when the computer assist is used by the readers. Analysis in terms of the alternative splitting shows large decreases in the corresponding modality-interaction components.

Conclusion. A solution to the problem of multivariate ROC analysis without the assumption of equal variance structure across modalities has been provided. Alternative formulations lead to consistent results related by a one-to-one mapping. A surprising result is that estimates of confidence intervals and numbers of cases and readers required for a specified confidence interval remain the same in the more general model as in the restricted model.

Key Words. Receiver operating characteristic (ROC); components-of-variance; jackknife; bootstrap.

Acad Radiol 2001; 8:605–615

[1] From the Offices of Science and Technology (HFZ-142) (S.V.B., R.F.W.) and Surveillance and Biometrics (HFZ-542) (G.C.), Center for Devices and Radiological Health, Food and Drug Administration, Rockville, MD 20857; and the Department of Radiology, Rossmann Laboratories, University of Chicago, Ill (C.E.M., Y.J.). Received July 24, 2000; revision requested December 14; revision received March 4, 2001; accepted March 5. S.V.B. supported in part by an appointment to the Research Participation Program at the Center for Devices and Radiological Health administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration. C.E.M. supported by grant R01-GM57622 from the National Institutes of Health and by a University of Chicago contract with the University of Iowa (Donald D. Dorfman, PI) under grant R01-CA62362 from the National Institutes of Health. Y.J. supported by U.S. Army Medical Research and Materiel Command grant DAMD17-00-1-0197. **Address correspondence to R.F.W.**

The field of random-effects receiver operating characteristic (ROC) analysis has made important advances during the past decade. Its major applications include the assessment of modalities for diagnostic imaging and computer-assisted diagnosis (CAD) and the comparison of competing diagnostic modalities. A particularly important paradigm is the multiple-reader, multiple-case (MRMC) approach in which every reader reads every patient case. This is the so-called reader study that allows for a proper accounting of both reader and case variance and thus provides estimates of uncertainties of ROC parameters that are said to be "generalizable to a population of readers as well as to a population of cases." This paradigm was first modeled by Swets and Pickett in 1982 (1). Dorfman, Ber-

baum, and Metz (DBM) later provided a more flexible theoretical and also more practical solution to the MRMC problem (2). Their use of a general linear model together with "jackknife" resampling allowed the application of standard analysis-of-variance (ANOVA) techniques. Their approach and several alternatives were discussed at a 1993 symposium, and the proceedings were published in a supplement to this journal (3).

Beiden, Wagner, and Campbell (BWC) have recently provided a review of some of the issues in random-effects ROC analysis, together with an alternative solution to the MRMC problem (4). The BWC analysis includes not only the estimation of uncertainties in performance estimates in the MRMC paradigm but also a method to uniquely decompose these uncertainties into contributions in a components-of-variance model (2,4,5). These components are referred to as the "variance structure" of the problem and include the case variability, the reader variability, various interactions among cases, readers, and modalities and, finally, experimental replication error. The BWC alternative to previous solutions involves the analysis of a set of population experiments in terms of the model components. In any realistic clinical context, such population experiments are not possible. The practical solution is to replace the set of population experiments with the set of corresponding bootstrap resampling experiments on the available finite data set. This leads to a system of linear equations that may be solved for estimates of the components of variance (ie, the sources of randomness). In turn, one then obtains estimates of the confidence intervals of interest, as well as the ability to size a pivotal study from a pilot study.

In the previous work, we followed the model and assumption of DBM, namely, that the reader and case variances and their interaction for one modality are so similar to those for the other modality that they can be assumed to be equal. A central goal of CAD and other evolutions in imaging technology, however, is to create new modalities that will outperform older ones—in ways that include reducing the magnitude of these components of variance. A comparison of the performance of such new and older technologies will therefore require a more general model.

In the present article, we extend our previous work to the more general case of unequal variance structures across two modalities under comparison. We will show how to solve for estimates of the variance structure for this more general MRMC paradigm. In the next section, we present one formulation of a solution to this estimation problem. An alternative formulation is presented in the Appendix. Our analysis is applied to the study of Jiang et al (6), in which unaided readers of suspicious mammographic clusters of calcifications were compared with readers who used a CAD modality as an adjunct. In a companion article (7), we analyze the uncertainties in the estimates of the variance structure.

MATERIALS AND METHODS

Following DBM (2), we analyze the MRMC paradigm within the framework of a general multivariate linear model for ROC parameter estimates. We will use the ROC area parameter, A_z (dropping the z for simplicity), to exemplify the model; the model is nevertheless applicable to any other ROC model parameter or accuracy index. For completeness, we repeat the multivariate linear model for an ROC accuracy index, A, used by DBM:

$$A_{ijkn} = \mu_i + r_j + c_k + (mr)_{ij} + (mc)_{ik} + (rc)_{jk} + (mrc)_{ijk} + z_{ijkn}, \quad (1)$$

where i indicates a particular imaging modality, j denotes a particular image reader, k is a particular case sample, and n is a particular replication of the experiment. (The index for case sample, k, is included in this model because DBM studied jackknife pseudo-values.) The term μ_i represents the contribution of modality i to the expected value of the accuracy index, while the remaining terms are independent zero-mean random variables. The terms with a single index are the reader and case contributions to the variability, with variances σ_r^2 and σ_c^2, respectively. The terms with two subscripts represent the two-way interactions between modality and reader, modality and case, and reader and case, with variances σ_{mr}^2, σ_{mc}^2, and σ_{rc}^2, respectively. The term with three subscripts represents the three-way interaction among modality, reader, and case, with variance σ_{mrc}^2. The last term is a pure error term in experimental reproducibility, with variance σ_z^2. For the case where multiple-reader experiments are conducted but readers do not independently repeat their readings, the last two terms, with variances σ_{mrc}^2 and σ_z^2, are inseparable, and we combine them into a single term with variance σ_ϵ^2.

A major distinction in applications of this model is that between random and fixed factors. A random factor is one that—on replication of the experiment—is drawn independently from a specified population; a fixed factor is one that remains unchanged on replication. As written

in Equation (1), modalities are considered fixed factors, while readers and cases are random factors. Finally, we note that it will not be necessary for the present article to invoke assumptions of normality.

The model of Equation (1) assumes that the variance structure is homogeneous across modalities. Our present interest, however, is the case in which this structure changes across modalities. A parsimonious model for this case and the application where readers do not independently repeat their readings can be obtained by making the reader, case, and reader-by-case interaction terms a function of modality (i), respectively, $r_j(i)$, $c_k(i)$, $(rc)_{jk}(i)$. It can be written as

$$A_{ijkm} = \mu_i + r_j(i) + c_k(i) + (mr)_{ij} + (mc)_{ik} + (rc)_{jk}(i) + \epsilon_{ijk}. \quad (2)$$

The two-way interaction terms involving modality m and readers r (or cases c) carry information related to the reader (or case) correlation across modalities; they do not require generalization. (All else being equal, the interaction strength is higher when the correlation is lower, and vice versa.) However, for the case where readers independently repeat their readings, the three-way interaction term also would not be made a function of modality, but the final term in Equation (1) would be. An alternative formulation, described in the Appendix, generalizes Equation (1) in a different way.

The variances produced by any linear model, such as Equation (2), and that contribute to observations over repeated experiments depend on which factors are held fixed and which are sampled randomly from a population when a particular ROC experiment is repeated. In reference 4 we showed that, for the equal-variance model considered there, it is possible to perform six population experiments, chosen from the family of 32 considered by Roe and Metz (5), that would allow one to solve for the six variance components in Equation (1), combining the final two components as just described. In the present work, we extend this approach to solve for nine components in the new model, using nine equations.

We use the notation of Roe and Metz (5), where variables to the left of the vertical bar in the subscript of an accuracy index are random factors, while those to the right are fixed factors. For example, suppose we consider replications of the experiment where readers R as well as cases C are drawn randomly from the population but the modality M is a fixed factor. All six variance components for a given modality contribute to the observed variance in this experiment. This is stated by the following expression, which, for the case of two modalities, provides two equations:

$$\mathrm{var}(A_{RC|M}) = \sigma_r^2(M) + \sigma_c^2(M) + \sigma_{mc}^2 + \sigma_{mr}^2 + \sigma_{rc}^2(M) + \sigma_\epsilon^2. \quad (3)$$

When readers are also a fixed effect, the pure reader term and the modality-by-reader term do not contribute. That experiment and observed variance are given by

$$\mathrm{var}(A_{C|MR}) = \sigma_c^2(M) + \sigma_{mc}^2 + \sigma_{rc}^2(M) + \sigma_\epsilon^2, \quad (4)$$

which also provides two equations when two modalities are being studied.

An experiment that is generally of most interest is the one in which two fixed modalities, M and M', are compared in terms of the ROC performance estimates obtained from randomly drawn reader and case samples. The population variance that is observed in that experiment can be calculated after subtracting two equations of the form of Equation (2) above:

$$A_{jk|1} - A_{jk|2} = [\mu_1 - \mu_2] + [r_j(1) - r_j(2)] + [c_k(1) - c_k(2)] + [rc_{jk}(1) - rc_{jk}(2)] + [mr_{1j} - mr_{2j}] + [mc_{1k} - mc_{2k}] + [\epsilon_{jk|1} - \epsilon_{jk|2}]. \quad (5)$$

The first term in square brackets on the right-hand side is not a random variable, and so it contributes no variance. The variance of the next term in square brackets involves the correlation of $r_j(1)$ and $r_j(2)$. In the present model, we take these components to be different in magnitude but perfectly correlated, that is, $r_j(1) = \gamma_r r_j(2)$, where γ_r is a constant. (We treat the pure case and reader-by-case components similarly.) Thus,

$$\mathrm{var}[r_j(1) - r_j(2)] = [\sigma_r(1) - \sigma_r(2)]^2. \quad (6)$$

This approach is consistent with the interpretation that the reader component was originally not a function of modality for the equal-variance model of Equation (1) and thus could be thought of as perfectly correlated across modalities in this special case to which the present model degenerates. More generally, of course, the reader variation

may not be perfectly correlated across modalities. However, the flexibility to include arbitrary correlations of readers (or cases) across modalities is achieved in the general linear model as used here through the presence of the interaction terms. (For split-plot designs, however, where the readers [or cases] are drawn independently for the two modalities [8], the present model would be modified to set the reader [or case] correlation across modality to zero.)

By similar steps, the variance of the complete difference expressed by Equation (5) can then be written as

$$\begin{aligned}\mathrm{var}(A_{RC|M} - A_{RC|M'}) = 2(\sigma^2_{mr} + \sigma^2_{mc} + \sigma^2_{\epsilon}) \\ + (\sigma_r(M) - \sigma_r(M'))^2 \\ + (\sigma_c(M) - \sigma_c(M'))^2 \\ + (\sigma_{rc}(M) - \sigma_{rc}(M'))^2. \quad (7)\end{aligned}$$

One similarly obtains the following results for the other experiments that are required in order to solve for all of the variance components in this model:

$$\mathrm{var}(A_{C|RM} - A_{C|R'M}) = 2(\sigma^2_{rc}(M) + \sigma^2_{\epsilon}), \quad (8)$$

$$\begin{aligned}\mathrm{var}(A_{C|RM} - A_{C|RM'}) = 2(\sigma^2_{mc} + \sigma^2_{\epsilon}) \\ + (\sigma_c(M) - \sigma_c(M'))^2 \\ + (\sigma_{rc}(M) - \sigma_{rc}(M'))^2, \quad (9)\end{aligned}$$

$$\begin{aligned}\mathrm{var}(A_{C|RM} - A_{C|R'M'}) = \sigma^2_{rc}(M) + \sigma^2_{rc}(M') \\ + 2(\sigma^2_{mc} + \sigma^2_{\epsilon}) \\ + (\sigma_c(M) - \sigma_c(M'))^2. \quad (10)\end{aligned}$$

Notice that Equations (3), (4), and (8) each describe two independent experiments ($M = 1$ or 2). The system of nine equations represented by Equations (3), (4), and (7)–(10) then expresses nine observable variances as a multivariate quadratic equation in the square roots of nine variance components. These equations reduce to the linear expressions in our previous work (4) for the case where the variances are equal across modalities.

The left-hand sides of Equations (3), (4), and (7)–(10) are observables that are independent of any model. Thus, we may equate the right-hand sides just derived for the present model with the corresponding right-hand sides that follow from the model for the equal-variance case given in BWC (4). When necessary to distinguish between models, we shall use the presubscript A to refer to components in the BWC model of reference 4 and the presubscript B to refer to components in the model described up to this point in the present article. (Components of variance in the alternative formulation described in the Appendix will be denoted by a presubscript C. Otherwise in this article, the components will refer to the present model, model B.) For example, equating Equation (8) as written above to the corresponding version of this for the equal-variance case yields

$$[{}_B\sigma^2_{rc}(1) + {}_B\sigma^2_{rc}(2)]/2 + {}_B\sigma^2_{\epsilon} = {}_A\sigma^2_{rc} + {}_A\sigma^2_{\epsilon}. \quad (11)$$

The average on the left-hand side of this equation results from the fact that in BWC (4) the observable quantity was taken to be the average over the fixed effect M, and thus we average over the two equations implied by Equation (8).

By repeating this exercise with Equations (4) and (3), and taking differences with Equation (11), two additional expressions parallel to Equation (11) can be found: one in which all versions of σ^2_c replace the corresponding versions of σ^2_{rc} and all versions of σ^2_{mc} replace the corresponding versions of σ^2_{ϵ}, and another in which all versions of σ^2_r replace the corresponding versions of σ^2_{rc} and all versions of σ^2_{mr} replace the corresponding versions of σ^2_{ϵ}. The complete parallel of these expressions with Equation (11) becomes apparent on recalling that σ^2_{ϵ} includes σ^2_{mrc}.

Another set of relationships can be found by first performing a similar exercise on Equations (10) and (4). The difference of the two results yields

$${}_A\sigma^2_c = {}_B\sigma_c(1)\,{}_B\sigma_c(2). \quad (12)$$

Similar results follow for the components σ^2_r and σ^2_{rc}. These expressions will be useful as a check on the results below.

We now proceed as in our previous analysis (4) where, in practice, we replace a given population experiment with the corresponding bootstrap experiment. (Details of the statistical bootstrap are reviewed in reference 4, based on Efron [9] and Efron and Tibshirani [10].)

The nonlinear system, Equations (3), (4), and (7)–(10), can be solved for the unknown variance components by

Components of Variance in Equal- and Unequal-Variance Models (All Values × 10^{-4})

Variance Component	Equal Variance	Unequal Variance	
		Modality 1	Modality 2
c	7.31	7.66	6.98
r	7.78	17.84	3.39
rc	2.11	10.92	0.41
mc	4.43	4.42	...*
mr	7.72	4.48	...
mrc/ϵ	14.00	10.45	...

*Ellipses indicate no new parameter; these components do not split in the new model.

numerical iteration. We first write the variance components as a vector $\boldsymbol{\sigma}$, whose transpose, T, is

$$(\boldsymbol{\sigma})^T = (\sigma_c(1), \sigma_c(2), \sigma_r(1), \sigma_r(2), \sigma_{mc}, \sigma_{mr}, \sigma_{rc}(1), \sigma_{rc}(2), \sigma_\epsilon). \quad (13)$$

Each of the nine equations is then rearranged such that the vector $\boldsymbol{\sigma}$ is on the left-hand side of the system; the right-hand side is then the remaining nonlinear operation on $\boldsymbol{\sigma}$, which we call f. The system can then be written as

$$\boldsymbol{\sigma} = f(\boldsymbol{\sigma}). \quad (14)$$

We use a method of simple iteration to solve this system, with the initial estimate being taken to be the solution of the linear system that results when the two structures are equal. This system of quadratic equations can be shown to have only one physically meaningful solution set, and thus the problem is well defined.

RESULTS

Application to CAD

We use the study of CAD by Jiang et al (6) to exemplify this approach. These authors compared the performance of 10 radiologists—unaided versus with the aid of a computer-assist modality—reading mammograms from 104 patients with clustered microcalcifications. The truth state for these patients was established with biopsy (46 malignant, 58 benign cases). ROC analysis for individual readers and also their average performance within the MRMC paradigm and model of DBM were published in reference 6. Here we use the methods described above to solve for the components of variance in these MRMC experiments, both in terms of the previous model that assumes equal variance structure across modalities and in terms of the new model that does not make that assumption.

Components of Variance

In the Table, we present the components of variance according to our present analysis within the two models: the first model assuming equal variances across modalities and the second model assuming unequal variances. Here, the first modality (modality 1) refers to the combination of mammographic images and unaided readers; the second modality (modality 2) refers to the combination of mammographic images and readers aided by the computerized feature extraction, fusion, and rating of probability of cancer described in reference 6.

The following observations can be made from the Table. In the equal-variance model, the reader component of variance and the case component of variance have similar strengths. The patient component of variance, which can be interpreted as the range of case difficulty as represented by the finite sample, hardly changed when we went to the unequal-variance treatment. The reader component of variance, which can be interpreted as a range of reader ability, "splits" into two quite unequal components in the unequal-variance model. Without the assist of CAD, the reader component is now seen to be much greater than the case component; the addition of CAD is seen to reduce this component more than fivefold. The reader-by-case component also splits into two quite unequal components, with a more than 25-fold reduction after the addition of CAD. Larger values of this component imply that the range of sampled case difficulty depends on the particular reader (or by the symmetry of that component, that the range of reader skills depends on the case); smaller values imply less such dependence. Thus, we take this splitting to indicate that the addition of CAD in the study of reference 6 almost eliminated the dependence of the range of case difficulty on the particular reader in that study (or, symmetrically, that it almost eliminated the dependence of the range of reader skills on the case).

For later reference (companion article, reference 7) these results are shown graphically in Figure 1, together with error bars on the model results that represent ± 1 standard deviation. In the companion article (7), we provide the analysis of uncertainty in these results. In a few words, the error bars are obtained by using a resampling technique known as the jackknife-

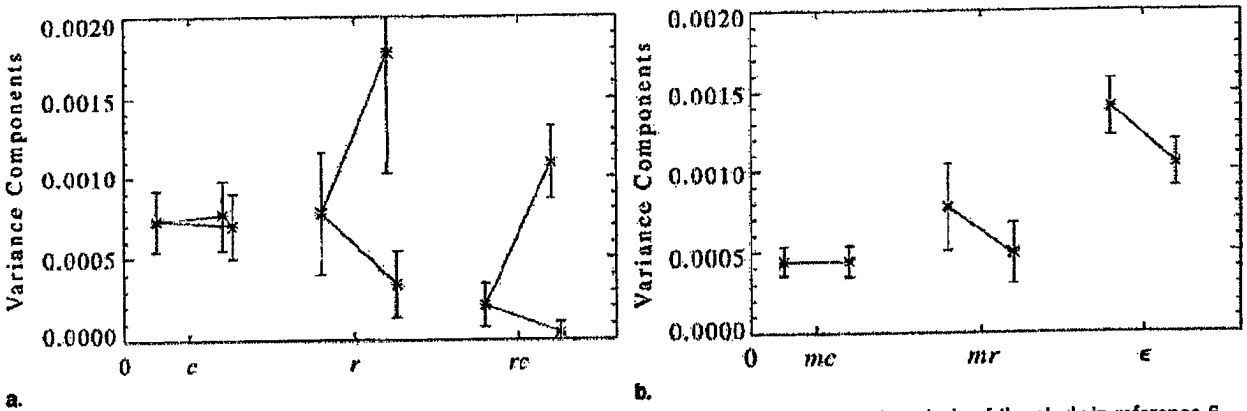

Figure 1. (a) Variance components *c* (case), *r* (reader), and *rc* (reader-by-case) in the present analysis of the study in reference 6. Vertical bars represent mean estimates, ± 1 standard deviation estimated with the method of reference 7. Unsplit components (to the left in each set of three) are estimated with the model of reference 4 (denoted model A). Splitting components (the pair to the right in each set) are estimated with model B of the present analysis (ie, not assuming equal variance structure across modalities). (b) Variance components *mc* (modality-by-case), *mr* (modality-by-reader), and ϵ (residual error) in the present analysis of the study in reference 6. These three components shift rather than split in going from model A of previous work to model B of the present article.

after-bootstrap (10), followed by linear propagation of variance for the model of equal variance structure or its modification for the model of unequal variance structure.

We note also that the model components in the unsplit model are indeed the geometric mean of the model components in the split model, consistent with the theoretical analysis of the model. Thus far, these observations are for the splitting components of the model. We now turn to the components that are not split in the new model.

In the new model, the reader-by-modality interaction changed to accommodate the new values of reader variance. There was little change in the case-by-modality interaction, as expected from the small change in the case component. Finally, the last component, or effective error term, is reduced when going to the new model. (The effective error term includes the contribution to the variance due to reader inconsistency that was called "jitter" in reference 11 and subsequent parlance.)

The shifts in the unsplit model terms that accompany the move to the more elaborate model can be accounted for by simple algebraic relationships. The change in the effective error term just noted, that is, the difference between the solution to the linear system, ${}_A\sigma^2_\epsilon$, and the corresponding solution to the quadratic system, ${}_B\sigma^2_\epsilon$, is simply related to the change in going from the solution to the linear system, ${}_A\sigma^2_{rc}$, to the solutions to the quadratic system, ${}_B\sigma^2_{rc}(1)$, ${}_B\sigma^2_{rc}(2)$. The relationship is found from Equations (11) and (12):

$$\begin{aligned}[{}_B\sigma^2_{rc}(1) + {}_B\sigma^2_{rc}(2)]/2 - {}_B\sigma_{rc}(1)\,{}_B\sigma_{rc}(2) \\ = [{}_B\sigma_{rc}(1) - {}_B\sigma_{rc}(2)]^2/2 \\ = {}_A\sigma^2_\epsilon - {}_B\sigma^2_\epsilon. \end{aligned} \quad (15)$$

That is, the shift in the nonsplitting component is the difference of the geometric mean and the arithmetic mean of the splitting components. Two additional expressions exactly parallel to Equation (15) can be found: one in which σ_c replaces σ_{rc} everywhere on the left-hand side and σ^2_{mc} replaces σ^2_ϵ on the right-hand side of Equation (15), and another in which σ_r replaces σ_{rc} everywhere on the left-hand side and σ^2_{mr} replaces σ^2_ϵ on the right-hand side of Equation (15). (Recall again that σ^2_ϵ contains σ^2_{mrc}.)

DISCUSSION

Inference and Experimental Design

An important consequence of the model and analysis above is that the present approach does not change the confidence intervals on the difference of ROC parameters between competing modalities, compared with our previous work (4). These confidence intervals are found from the single-bootstrap experiment represented by the left-hand side of Equation (7). The right-hand side of this

equation is new in the present model, and thus the interpretation is new. The input to the equation represented by the left-hand side has not changed, however.

The new model also does not change the design of a pivotal study from results of a pilot study that was described in our previous analysis (4). In that analysis, the variance components mc, mr, and ϵ were the only contributors to the estimation of the numbers of cases and readers required for a specified confidence interval on the difference of ROC parameters between competing modalities, but in the present analysis there are nine contributions to that estimation task (right-hand side of Equation [7]). Although the former three terms may be reduced in the new model, inspection of Equation (15) and its analogues shows that this reduction is exactly offset by the remaining terms of Equation (7). Thus, the design of experiments according to the previous model and model parameters obtained in reference 4 is unchanged, if only the difference in performance between two modalities is of interest. However, the partitioning of the variances obtained in the present work provides additional insight for the entire family of possible experiments embraced in Equations (3), (4), and (7)–(10).

Since confidence intervals on differences between modalities and associated inferences based on them in our analysis do not change when going to the new model, it would be reasonable to expect that inferences based on an elaboration of the DBM analysis to the case of unequal variance structures across modalities would also remain unchanged. We have argued in reference 4 that our analysis is a distribution-free generalization of the approach of DBM. Since inferences based on this generalization remain unchanged when the variance structure is allowed to change across modalities, inferences based on an elaboration of DBM (ie, use of the jackknife rather than the bootstrap) might also be expected to remain unchanged. In the Appendix, we present an alternative to model B that contains no expressions nonlinear in the variance components and could thus be readily incorporated into the method of DBM. We refer to this alternative as model C. In the approach of the present article, inferences and design of experiments based on model C are identical to those based on model B, and they are thus identical to those based on BWC (4) when estimation of differences between modalities is the object of the experiment.

Generality of the Present Work

An anonymous reviewer (December 2000) has suggested that one could address the present problem by a natural extension of the DBM approach, using jackknifed pseudo-values with the PROC MIXED routines in the SAS software package (12). This will lead not only to estimates of the confidence intervals of interest but also to maximum-likelihood (ML) or residual (often called restricted) maximum-likelihood (REML) estimates of the variance components. The approach of using REML to obtain estimates of the variance components had also been mentioned to one of us (R.F.W.) previously (D.D. Dorfman, oral communication, 1999). We agree that this is indeed a reasonable alternative to the present approach, but it does not address the level of generality we seek here. We summarize this issue as follows.

The BWC approach (4), and its extension to the case of unequal variance structures as provided above, is built on the same general components-of-variance model used by DBM. However, it replaces the jackknife and ANOVA with a family of bootstrap resampling experiments and a corresponding system of equations that lead to explicit solutions for the variance components and confidence intervals of interest. It is thus a distribution-free approach, whereas classic ANOVA is based on the assumption of normality for all the components. (REML also requires assumptions for the relevant distributions.)

An additional feature of the present approach was cited in reference 4. The bootstrap includes not only the leave-one-out jackknife, but also more general leave-X-out terms where X is greater than one, among the other kinds of terms that sampling with replacement generates. For the case where the statistic of interest is linear, all of the terms that can contribute to the calculation of that statistic on a single-bootstrap pass are already included in the jackknifed data sets; this is not true of nonlinear statistics, that is, statistics that involve interactions between the data points two or more at a time (10). The nonparametric estimate of ROC area, for example, includes sums of rankings of data points two at a time (13,14) and thus falls into the latter category. Thus, the leave-one-out jackknife does not in general capture all of the information in the data regarding this statistic. Nevertheless, only small differences were found in reference 4 between the DBM and BWC methods for the variance structures and samples sizes studied there. Also, in our (unpublished) Monte Carlo simulations of bootstrap and jackknife estimates of variance for the nonparametric measure of ROC area, small differences between mean estimates were seen, but only when the number of patients per class was smaller than 25. This issue bears further investigation, including the case of parametric accuracy measures.

Finally, we emphasize a general point about the philosophy of the bootstrap made by Efron and Tibshirani (10). The empirical distribution function is the nonparametric ML estimate of the population distribution. In this sense, the nonparametric bootstrap provides "nonparametric ML estimates" in the language of reference 10, or "distribution-free" ML estimates in language that we and others prefer. The system of equations used here to propagate those estimates back into estimates of the variance components will thus also lead to distribution-free ML estimates. (This follows since the ML estimate of a function of a parameter of interest is that function of the ML estimate of the parameter.) For all of the above points, we would argue that the approach of reference 4 and its present extensions are the most general proposed so far for the family of problems under consideration here.

CONCLUSIONS

The present approach to random-effects ROC analysis extends our previous work (4) to the case where the variance structure may change across modalities. An example comparing unaided readers with readers assisted by CAD showed that both the reader and the reader-by-case components of variance were greatly reduced after the addition of CAD. These results are consistent with previous expectations regarding that study (15), but such results had not been previously isolated quantitatively.

Several comments regarding the future are in order. The present model provides a quantitative framework for interpreting the variability in MRMC studies in terms of a model of the components of that variability. It may thus offer the opportunity to contribute to the solution of several outstanding problems in the field of medical image science. The first of these is the connection between physical performance measurements on diagnostic imaging systems, that is, measurements of "image quality," and measures of clinical outcome such as the ROC curve (16,17). The variability observed at present in mammographic imaging (18), to take just one example, may mask the gains to be expected from evolution of the physical performance of mammographic imaging systems. The present approach may make it possible to peel back this mask with an efficient clinical experimental design.

The ability to isolate the contributions to variability in performance that arise from the reader from those that arise from the patient and the imaging system opens up new possibilities for imaging system optimization. The professional community of radiologists may be better able to quantitatively measure and fine-tune their training of readers, while the professional community of physicists and engineers of imaging systems may be better able to fine-tune their system designs, each with the appropriate focus and emphasis.

Finally, the emergence of the field of computer-assisted reading of images adds another layer of complexity to the problem of assessing diagnostic imaging modalities. The present work may contribute toward extending our understanding and optimization of the interface between the imaging physics and human image readers to the further interface of these with computerized reading-assist modalities.

APPENDIX

The formulation described in the body of the present article models the situation where the variance structure is allowed to be unequal across modalities by splitting the case, reader, and reader-by-case components in the general linear model, that is, it makes them a function of modality; it leaves the modality-by-case, modality-by-reader, and modality-by-reader-by-case components unsplit. An alternative to this model can be constructed by splitting the latter three components and leaving the former three components unsplit. We present the alternative model equations here, together with a demonstration of a one-to-one correspondence between the two alternative models.

In the alternative model, the modality-by-case, modality-by-reader, and modality-by-case-by-reader terms are functions of modality, i, and are written $mc_{ij}(i)$, $mr_{ik}(i)$, and $(mrc)_{ijk}(i)$, respectively. These components are taken to be independent across modalities. The linear model of Equation (1) then becomes

$$A_{ijkn} = \mu_i + r_j + c_k + (rc)_{jk} + (mr)_{ij}(i) + (mc)_{ik}(i) + (mrc)_{ijk}(i) + z_{ijkn}(i). \quad \text{(A1)}$$

The strengths of the components of variance of this model will be distinguished from those of the models discussed in the body of the present article by the addition of a presubscript C. Here, as earlier, we consider the case of no replication and thus set ${}_C\sigma^2_\epsilon(i) = {}_C\sigma^2_{mrc}(i) + {}_C\sigma^2_z(i)$. Equations (3), (4), and (7)–(10) for the observable variances in terms of the model components of variance for the case of no replication then become

$$\mathrm{var}(A_{RC|M}) = {}_C\sigma^2_r + {}_C\sigma^2_c + {}_C\sigma^2_{mc}(M) + {}_C\sigma^2_{mr}(M) + {}_C\sigma^2_{rc} + {}_C\sigma^2_\epsilon(M), \quad \text{(A2)}$$

$$\mathrm{var}(A_{C|MR}) = {}_C\sigma_c^2 + {}_C\sigma_{mc}^2(M) + {}_C\sigma_{rc}^2 + {}_C\sigma_\epsilon^2(M), \quad \text{(A3)}$$

$$\mathrm{var}(A_{RC|M} - A_{RC|M'}) = {}_C\sigma_{mr}^2(1) + {}_C\sigma_{mr}^2(2) + {}_C\sigma_{mc}^2(1) + {}_C\sigma_{mc}^2(2) + {}_C\sigma_\epsilon^2(1) + {}_C\sigma_\epsilon^2(2), \quad \text{(A4)}$$

$$\mathrm{var}(A_{C|RM} - A_{C|R'M}) = 2{}_C\sigma_{rc}^2 + 2{}_C\sigma_\epsilon^2(M), \quad \text{(A5)}$$

$$\mathrm{var}(A_{C|RM} - A_{C|RM'}) = {}_C\sigma_{mc}^2(1) + {}_C\sigma_{mc}^2(2) + {}_C\sigma_\epsilon^2(1) + {}_C\sigma_\epsilon^2(2), \quad \text{(A6)}$$

$$\mathrm{var}(A_{C|RM} - A_{C|R'M'}) = 2{}_C\sigma_{rc}^2 + {}_C\sigma_{mc}^2(1) + {}_C\sigma_{mc}^2(2) + {}_C\sigma_\epsilon^2(1) + {}_C\sigma_\epsilon^2(2), \quad \text{(A7)}$$

As with Equations (3), (4), and (7)–(10), these equations also reduce to the expressions in our previous work (4) for the case where the variances are equal across modalities. Notice, however, that Equations (A2)–(A7) are now linear in the model components. Thus, they may be solved for these components by linear algebra in the same manner as was used in our earlier work (4).

As noted earlier, the left-hand sides of Equations (A2)–(A7) are observables that are independent of any model. Thus, we may equate the right-hand sides for the present model with the corresponding right-hand sides of Equations (3), (4), and (7)–(10) to discover the relations between the components of variance in the two models. For example, equating the right-hand sides of Equation (8) and Equation (A5) yields

$${}_B\sigma_{rc}^2(M) + {}_B\sigma_\epsilon^2 = {}_C\sigma_{rc}^2 + {}_C\sigma_\epsilon^2(M), \quad \text{(A8)}$$

where, as above, the presubscript B refers to the model in the body of the present article. Similarly, equating the right-hand sides of Equation (4) and Equation (A3), and subtracting Equation (A8) yields

$${}_B\sigma_c^2(M) + {}_B\sigma_{mc}^2 = {}_C\sigma_c^2 + {}_C\sigma_{mc}^2(M), \quad \text{(A9)}$$

and equating the right-hand sides of Equation (3) and Equation (A2) and subtracting the results in Equations (A8) and (A9) yields

$${}_B\sigma_r^2(M) + {}_B\sigma_{mr}^2 = {}_C\sigma_r^2 + {}_C\sigma_{mr}^2(M). \quad \text{(A10)}$$

The complete parallelism of Equations (A8)–(A10) may be more apparent on recalling that σ_ϵ^2 in all models here contains σ_{mrc}^2. These three equations show that changing from model B to model C changes the distribution of variance strength *within* the three compartments defined by these three equations, but it does not redistribute variance strength *across* these three compartments or equations. Continuing in this way, we may solve for the components of model B in terms of those of model C and vice versa, as we now show.

Equating the right-hand sides of Equations (10) and (A7), equating the right-hand sides of Equations (4) and (A3), and subtracting yields

$${}_C\sigma_c^2 = {}_B\sigma_c(1){}_B\sigma_c(2). \quad \text{(A11)}$$

Finally, the equivalence of the right-hand sides of Equations (9) and (A6), and of Equations (7) and (A4), leads in a similar way to

$${}_C\sigma_{rc}^2 = {}_B\sigma_{rc}(1){}_B\sigma_{rc}(2) \quad \text{(A12)}$$

and

$${}_C\sigma_r^2 = {}_B\sigma_r(1){}_B\sigma_r(2). \quad \text{(A13)}$$

Thus, from Equations (A8)–(A10) and Equations (A11)–(A13), we have also

$${}_C\sigma_{mr}^2(M) = {}_B\sigma_r^2(M) + {}_B\sigma_{mr}^2 - {}_B\sigma_r(1){}_B\sigma_r(2),$$

$${}_C\sigma_{mc}^2(M) = {}_B\sigma_c^2(M) + {}_B\sigma_{mc}^2 - {}_B\sigma_c(1){}_B\sigma_c(2),$$

$${}_C\sigma_\epsilon^2(M) = {}_B\sigma_{rc}^2(M) + {}_B\sigma_\epsilon^2 - {}_B\sigma_{rc}(1){}_B\sigma_{rc}(2). \quad \text{(A14)}$$

Equations (A12)–(A14) express the components of model C in terms of the components of model B. The relationships in the other direction may be obtained as follows.

The first equation of the set, Equation (A14), provides two equations whose difference is

$${}_B\sigma_r^2(1) - {}_B\sigma_r^2(2) = {}_C\sigma_{mr}^2(1) - {}_C\sigma_{mr}^2(2). \quad \text{(A15)}$$

The square of Equation (A13) may be used to rewrite the second term of Equation (A15) in terms of the first (and vice versa), providing a quadratic equation in ${}_B\sigma_r^2(1)$ (or

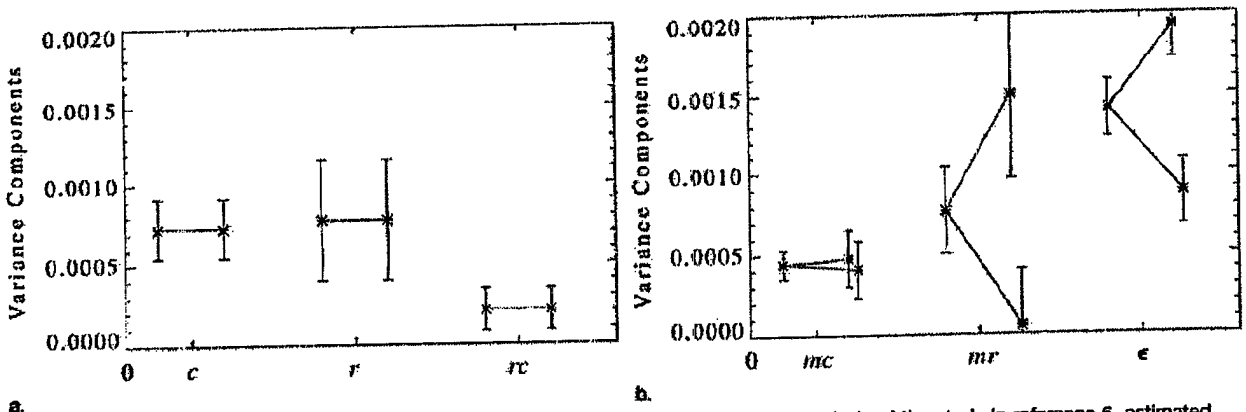

Figure A1. (a) Variance components c (case), *r* (reader), and *rc* (reader-by-case) in the analysis of the study in reference 6, estimated with model A of reference 4 (vertical bars to the left in each pair) and model C of the Appendix (vertical bars to the right in each pair). Vertical bars represent mean estimates, ± 1 standard deviation obtained with the methods of reference 7. Note that these three components remain unchanged in going from model A to model C. (b) Variance components *mc* (modality-by-case), *mr* (modality-by-reader), and ϵ (residual error) in the analysis of the study in reference 6. Vertical bars represent mean estimates, ± 1 standard deviation estimated with the method in reference 7. Unsplit components (to the left in each set of three) are estimated with the model of reference 4 (denoted model A in the present article). Splitting components (the pair to the right in each set) are estimated with model C of the Appendix.

${}_B\sigma_r^2(2)$). The solutions are

$$ {}_B\sigma_r^2(1) = [({}_C\sigma_r^2)^2 + (b/2)^2]^{1/2} + b/2, \qquad \text{(A16)} $$

$$ {}_B\sigma_r^2(2) = [({}_C\sigma_r^2)^2 + (b/2)^2]^{1/2} - b/2, $$

where

$$ b = {}_C\sigma_{mr}^2(1) - {}_C\sigma_{mr}^2(2). $$

The form of Equation (A16) shows that there is only one nonnegative solution. Parallel solutions of identical form can be found in the same way for ${}_B\sigma_c^2(M)$ and ${}_B\sigma_{rc}^2(M)$.

Finally, expressions for the nonsplitting components in model B may be obtained in terms of the components in model C by combining Equation (A16) and its analogs with Equations (A8)–(A10).

The present exercise demonstrates a one-to-one mapping between model C and model B. The selection between them thus appears to be a matter of intuitive appeal or taste. An appealing feature of model B is that the components that are split correspond to populations (cases, readers, readers-by-cases) that seem intuitively natural, and thus model B is pedagogically attractive. On the other hand, the splitting employed by model C may be more intuitive for some, and an attractive feature of this model is the fact that the equations to which it leads, Equations (A2)–(A7), remain linear in a set of independent variance components. As a consequence, it is suitable for incorporation into conventional ANOVA (as in DBM [2], for example). (The feature of linearity is not an issue for the multiple-bootstrap approach; the choice of model B versus model C leads to only small differences in the computer coding that is required in that approach.) Finally, model C requires no adjustment to accommodate split-plot designs.

In the same manner as above, we may also show that the interaction components in model A are related to those in model C as simple arithmetic averages:

$$ {}_A\sigma_{mr}^2 = [{}_C\sigma_{mr}^2(1) + {}_C\sigma_{mr}^2(2)]/2, \qquad \text{(A17)} $$

and similarly for σ_{mc}^2 and σ_ϵ^2. Now, it is Equation (A4) that determines the confidence intervals on the difference of ROC accuracy measures across two fixed modalities when readers and cases are taken as random effects. The left-hand side of Equation (A4) describes the underlying population or bootstrap experiment. The right-hand side is its decomposition according to model C and is proportional to the sum of three averages, namely, the right-hand side of Equation (A17) and the analogous terms for the σ_{mc}^2 and σ_ϵ^2 components. The averaged components are precisely the terms that contribute in model A, the equal-variance model. Thus, as far as the confidence interval of interest here is concerned, no new issues arise when moving from model A to model C. (This is the same conclu-

sion found in the body of the article when moving from model A to model B.)

The example of the present article may be analyzed in terms of model C of this Appendix. The results are shown in Figure A1a and A1b. The particular details of these figures are different from those in Figure 1a and 1b of the text, because the absolute levels of the quantities that are split differ across the two models. However, because of the one-to-one correspondence between the two models, there is no fundamental difference between the conclusions drawn from either set of figures.

DEDICATION

The authors dedicate this work to the memory of Donald D. Dorfman, PhD, of the University of Iowa, who passed away on April 15, 2001. Don's singular contributions to this field have always been an inspiration to the present authors. The field will not be the same without him.

REFERENCES

1. Swets JA, Pickett RM. Evaluation of diagnostic systems: methods from signal detection theory. New York, NY: Academic Press, 1982.
2. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol 1992; 27:723–731.
3. Gatsonis CA, Begg CB, Wieand S, eds. Advances in statistical methods for diagnostic radiology: a symposium. Acad Radiol 1995; 2(suppl 1):S1–S84.
4. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative methodology for random-effects ROC analysis. Acad Radiol 2000; 7:341–349.
5. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Acad Radiol 1997; 4:587–600.
6. Jiang Y, Nishikawa RM, Schmidt RA, Metz CE, Giger ML, Doi K. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol 1999; 6:22–33.
7. Beiden SV, Wagner RF, Campbell G, Chan HP. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. Acad Radiol 2001; 8:616–622.
8. Dorfman DD, Berbaum KS, Lenth RV, Chen YF. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: split plot experimental design. Proc SPIE 1999; 3663:91–99.
9. Efron B. The jackknife, the bootstrap and other resampling plans. Philadelphia, Pa: Society for Industrial and Applied Mathematics, 1982.
10. Efron B, Tibshirani RJ. An introduction to the bootstrap: monographs on statistics and applied probability. New York, NY: Chapman & Hall, 1993.
11. Goodenough DJ, Metz CE. Implications of a "noisy" observer to data processing techniques. In: Raynaud C, Todd-Pokropek A, eds. Information processing in scintigraphy. Orsay, France: Commisseriat a l'Energie Atomique, Departement de Biologie, Service Hospitalier Frederic Joliot, 1975.
12. SAS Institute. SAS/STAT software: changes and enhancements through release 6.12. Cary, NC: SAS Institute, 1997; 573–577.
13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–845.
14. Campbell G, Douglas MA, Bailey JJ. Nonparametric comparison of two tests of cardiac function on the same patient population using the entire ROC curve. In: Ripley KL, Murray A, eds. Computers in cardiology. Long Beach, Calif: IEEE Computer Society, 1989; 267–270.
15. Jiang Y, Nishikawa RM, Schmidt RA, Toledano AY, Doi K. The potential of computer-aided diagnosis (CAD) to reduce variability in radiologists' interpretation of mammograms. Radiology (in press).
16. International Commission for Radiation Units and Measurements. Medical imaging: the assessment of image quality. ICRU Report no. 54. Bethesda, Md: International Commission for Radiation Units and Measurements, 1996.
17. Metz CE, Wagner RF, Doi K, Brown DG, Nishikawa RM, Myers KJ. Toward consensus on quantitative assessment of medical imaging systems. Med Phys 1995; 22:1057–1061.
18. Beam C, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med 1996; 156:209–213.

An optimal method for combining two correlated diagnostic assessments with application to computer-aided diagnosis

Yulei Jiang and Charles E. Metz
Kurt Rossmann Laboratories for Radiologic Image Research
Department of Radiology, The University of Chicago, Chicago, IL 60637

ABSTRACT

We are developing computer-aided diagnosis (CAD) methods that produce a quantitative diagnostic assessment, such as the likelihood of malignancy of a breast lesion. Radiologists who use this computer aid must combine the computer's quantitative assessment with their own. No theoretical or empirical methods are currently available to help radiologists perform this task. Results of recent observer studies show that while CAD helps radiologists improve performance, radiologists' *ad hoc* performance tends to be inferior to that of the computer alone, indicating that they are unable to use computer aids optimally. We have developed a general method to combine two correlated diagnostic assessments. We calculate a likelihood ratio based on a bivariate binormal model that describes the joint probability density of the latent decision variables from two diagnostic assessments. To the extent that the bivariate binormal model is valid and that the model's parameters can be estimated reliably, results that we obtain in this way will be optimal because that likelihood ratio is used by the ideal observer in combining the diagnostic assessments. Preliminary results indicate that this method can produce better performance than that achieved by radiologists when they use computer aids in an *ad hoc* way. This method can potentially help radiologists use quantitative computed diagnostic assessments optimally, thereby surpassing the computer in accuracy.

Keywords: computer-aided diagnosis (CAD), receiver operating characteristic (ROC) analysis, observer performance, ideal observer approximation.

INTRODUCTION

Computer-aided diagnosis (CAD) involves either a binary or a quantitative form of computer aid. Computer-aided *detection* usually involves a binary-form computer aid, such as an arrow that indicates the location of a potential lesion in a mammogram. On the other hand, computer-aided *diagnosis* of malignant and benign lesions often involves a quantitative computer aid[1-4]. For example, the computer aid may be an estimate of the likelihood of malignancy[2]. When a quantitative computer aid is involved, the computer's quantitative assessment must be combined in some way with a radiologist's diagnostic assessment. In published studies, radiologists are responsible for this task of combining two sources of diagnostic assessments[1-4]. Radiologists do so in an *ad hoc* way because there is no formal methods for radiologists to perform this task. Results of the published studies show that radiologists are able to improve their diagnostic performance by using a computer aid but sometimes cannot perform as well as the computer[2, 3]. This inability of radiologists to outperform the computer indicates that they are not always able to combine a quantitative computer aid with their own diagnostic assessments optimally.

Our purpose was to develop a generally "optimal" method for combining two sources of correlated diagnostic assessments and to apply this method to CAD. If such a method can be developed, then radiologists may be able to improve their performance in CAD more than that achievable from *ad hoc* use of the computer aid. Our method was based on the calculation of a likelihood ratio that takes into account the individual accuracies of the two sources of diagnostic assessments as well as their correlation. Because the ideal observer also uses this likelihood ratio, our method is theoretically optimal if the model that was used to calculate the likelihood ratio is valid and if the model parameters can be estimated reliably.

Image Perception and Performance, Elizabeth A. Krupinski,
Dev P. Chakraborty, Editors, Proceedings of SPIE Vol. 4324 (2001)

THORETICAL BACKGROUND

The bivariate binormal model and the likelihood ratio

The bivariate binormal model, illustrated schematically in Fig. 1, was developed by Metz *et al.* for testing the significance of differences between ROC curves measured from correlated data[5]. The marginal distributions of this bivariate binormal model reduce to a pair of conventional univariate binormal models that represent the diagnostic accuracies of the radiologist and of the computer. The shapes of the ellipsoids in the bivariate binormal model are determined by the individual accuracies of the radiologist and the computer and by the correlation between their latent decision variables. This model is appropriate for our present purpose because it takes into account the individual accuracies of the radiologist and the computer and the correlation of their diagnostic assessments.

Based on the bivariate binormal model, we define a likelihood ratio

$$LR \equiv \frac{\text{Prob}(\vec{x} \mid \text{cancer})}{\text{Prob}(\vec{x} \mid \text{benign})}$$

We will use this likelihood ratio (or equivalently, the logarithm of this likelihood ratio) as the combined diagnostic assessment of the radiologist and the computer. Because the ideal observer also uses this likelihood ratio to combine the diagnostic assessments of the radiologist and the computer, our method is theoretically optimal. However, in practice, its results will depend on the validity of the bivariate binormal model in a particular situation and on whether the model parameters can be estimated reliably. The log-likelihood ratio is a quadratic function of the latent decision variables used by the radiologist and the computer because the bivariate binormal model involves a generally quadratic function of the same latent decision variables[5]. Therefore, we call this method "quadratic averaging."

Quadratic averaging requires a "training" process to estimate the bivariate-binormal model parameters. In general, this requires a set of "training" cases comprised of the diagnostic assessments of a radiologist and a computer as well as the diagnostic "truth" in all cases. Metz's CLABROC algorithm can be used to estimate the bivariate-binormal model parameters from these training cases[5]. After the model parameters are determined, quadratic averaging can be applied to new cases.

Arithmetic averaging

Metz and Shen described a method of unweighted arithmetic averaging of diagnostic assessments[6]. They showed that arithmetic averaging improves the resulting ROC curve by suppressing reader variations that are associated with the

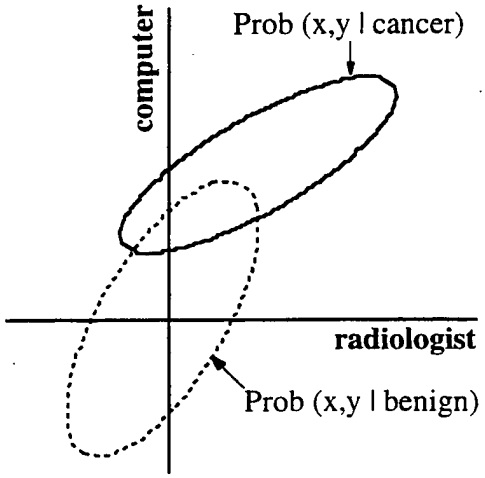

Figure 1. Schematic illustration of the bivariate binormal model

diagnostic assessments. Swensson *et al.* described a similar method that uses the median for the same task[7]. These methods are attractive because they do not require a "training" process for the estimation of any model parameters and because they are straightforward to implement in practice. We compare the results of quadratic averaging and of arithmetic averaging.

MATERIALS AND METHODS

An observer-study dataset

We used data from a recent observer study[2]. The purpose of that study was to compare radiologists' performance in differentiating malignant and benign clustered microcalcifications in mammograms with and without the aid of a computer-estimated likelihood of malignancy. The observer study consisted of 104 cases of mammograms (46 contained cancers and 58 contained benign lesions) and 10 radiologists who read all cases both without and with the computer aid. The study design followed the multiple-reader, multiple-case (MRMC) paradigm[8]. The details of the observer study are described elsewhere[2].

In the observer study, each radiologist provided his or her diagnostic confidence that a lesion was malignant for all cases when reading the mammograms without the computer aid. In addition, each radiologist also provided his or her diagnostic confidence when the computer-estimated likelihood of malignancy was available to the radiologist. Finally, the computer provided an estimate of the likelihood of malignancy in all cases. Both radiologists' diagnostic confidence ratings and the computer-estimated likelihood of malignancy were on a continuous scale of 0-100%.

Combining radiologists' unaided diagnostic assessments and the computer-estimated likelihood of malignancy

Arithmetic averaging and quadratic averaging were used to combine the radiologists' unaided diagnostic assessments with the computer's estimate of the likelihood of malignancy. For arithmetic averaging, a radiologist's diagnostic confidence rating (0-100%) and the computer-estimated likelihood of malignancy (0-100%) were simply averaged arithmetically to produce a combined diagnostic score.

Quadratic averaging was done in two ways. First, the bivariate-binormal model parameters were estimated from all available cases and were then used to produce a combined diagnostic score for the same cases. Because this method is equivalent to a re-substitution plan[9, 10], we refer to this implementation of quadratic averaging as Quad-RS. Second, the bivariate-binormal model parameters were estimated from all cases except one case and the resulting model parameters were used to produce a combined diagnostic score for the one left-out case. Therefore, for each case, the model parameters were re-estimated from a different set of cases. Because this method is equivalent to a leave-one-out re-sampling plan[9, 10], we refer to it as Quad-LOO.

Computer simulations

To confirm the results from actually pairing each of the 10 radiologists with the computer, we performed computer simulations using the same bivariate-binormal model parameters as those from actually pairing the radiologists and the computer. In the simulations, we generated a "training" dataset and a separate "test" dataset by random sampling from the assumed bivariate binormal models. The "training" dataset was used to estimate the bivariate-binormal model parameters required for quadratic averaging; these model parameters were then used to apply quadratic averaging on the "test" cases. Arithmetic averaging was applied to the "test" cases only and the "training" dataset was not used for arithmetic averaging. The numbers of cases in both the training and the test datasets were the same as in the observer study: 46 malignant and 58 benign cases. The simulations consisted of a total of 1,000 repetitions.

RESULTS

Results obtained from the observer study

We summarize the results of the observer study that are relevant to a comparison between the results of arithmetic and quadratic averaging. A complete report of the observer study can be found elsewhere[2]. In this observer study, the 10 radiologists achieved an average A_z value of 0.61 when they read the mammograms without the computer aid and an average A_z value of 0.75 when the computer-estimated likelihood of malignancy was available to them. The improvement in A_z was statistically significant ($p < 0.0001$; Dorfman-Berbaum-Metz method[8]). However, the performance of the computer alone was higher than the performance of the radiologists either without or with the computer aid. The A_z value of the computer alone was 0.80 and the differences between this computer performance and the performance of the radiologists were statistically significant ($p < 0.0001$ without and $p = 0.002$ with the computer aid; Student's t-test for paired data). Figure 2 shows (solid lines) the A_z values of each radiologist's unaided and computer-aided performance as well as the performance of the computer. (For comparison purpose, the computer's A_z value is replicated over each radiologist's data and plotted as a straight line.)

Results obtained from combining radiologists' unaided diagnostic assessments and the computer's estimate of the likelihood of malignancy

Results of arithmetic averaging are shown in Fig. 2 (long-dash line). The average A_z value obtained from arithmetic averaging was 0.76. For 5 readers, the arithmetic-average A_z values were higher than the A_z values achieved by the radiologists when they actually had the computer aid. For the other 5 readers, the opposite was true. For all radiologists except one, the arithmetic-average A_z values were lower than the A_z value of the computer alone.

Results of quadratic averaging are also shown in Fig. 2. The Quad-RS results (dotted line)—when the parameters of the bivariate binormal model were estimated from the same cases as those that produced the quadratic averaging results—

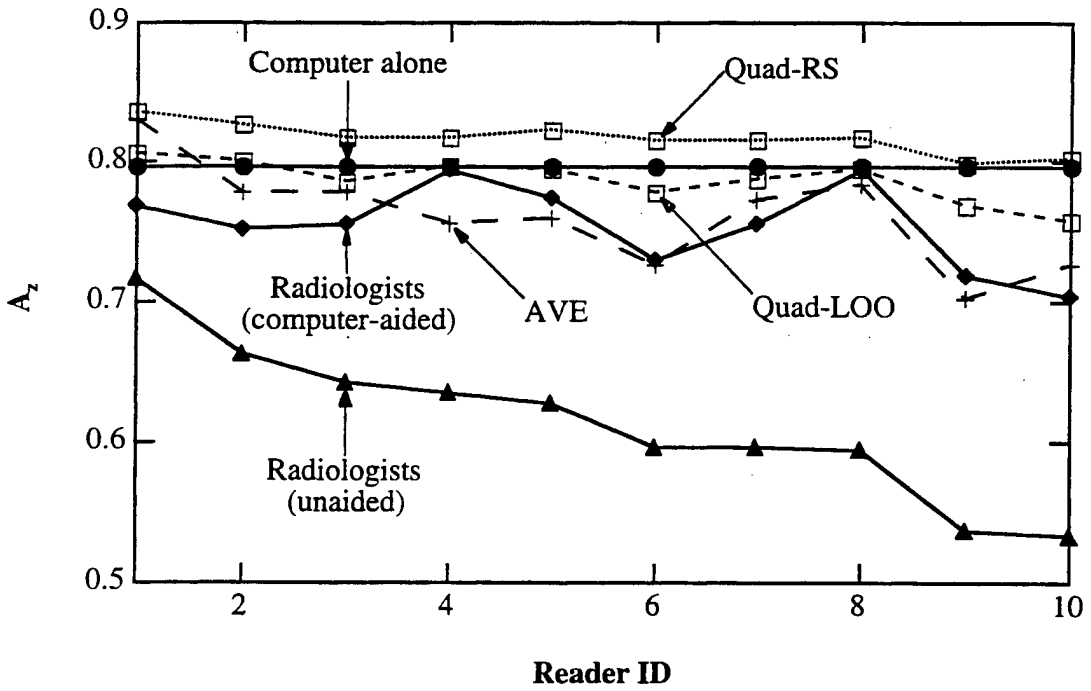

Figure 2. Comparison of diagnostic performance of 10 radiologists and a computer obtained in the observer study and three methods of combining the radiologists' unaided diagnostic confidence ratings with the computer-estimated likelihood of malignancy.

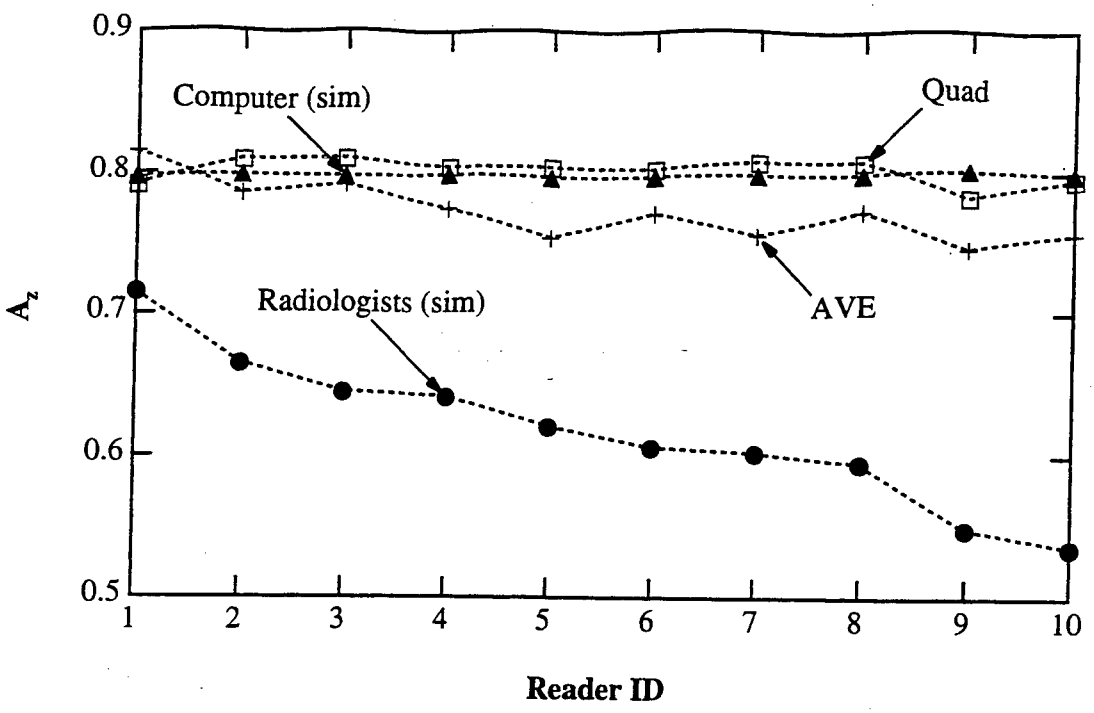

Figure 3. Simulation results.

show an average A_z value of 0.82. For all 10 readers, the Quad-RS A_z values were higher than the A_z values achieved by the radiologists when they actually had the computer aid. In addition, for all 10 readers, the Quad-RS A_z values were higher than the A_z value of the computer alone. Finally, for all 10 readers, the Quad-RS A_z values were higher than the arithmetic-average A_z values.

Results of Quad-LOO (short-dash line)—when parameters of the bivariate binormal model were estimated from cases that are different from those that produced the quadratic averaging results—show an average A_z value of 0.79. For all 10 readers, the Quad-LOO A_z values were equal to or higher than the A_z values achieved by the radiologists when they actually employed the computer aid. In addition, for 5 radiologists, the Quad-LOO A_z values were equal to or higher than the A_z values of the computer alone. Finally, for all radiologists except one, the Quad-LOO A_z values were higher than the A_z values of arithmetic averaging.

Simulation results

Simulation results are shown in Fig. 3. These results generally confirm the results of actually pairing each of the radiologists with the computer. In the simulations, the average A_z value of arithmetic averaging was 0.77. Except for one radiologist, the arithmetic-average A_z values were lower than the A_z values of the computer alone. The average A_z value of quadratic averaging was 0.80. Except for one radiologist, the quadratic-average A_z values were higher than the arithmetic-average A_z values. In addition, for 7 radiologists, the quadratic-average A_z values were higher than the A_z values of the computer alone.

DISCUSSION

Quadratic averaging is a theoretically optimal method for combining two sources of correlated diagnostic assessments. This method can be applied in CAD to combine quantitative diagnostic assessments of a radiologist and of a computer. Although practical limitations apply, our study shows that quadratic averaging can consistently produce results that are better than those achieved by radiologists who use a quantitative computer aid in an *ad hoc* way. In addition,

quadratic averaging can consistently produce results that are better than those of arithmetic averaging, at least in situations where the individual accuracies of the diagnostic assessments to be combined differ substantially and/or the two conditional distributions of the bivariate binormal model differ substantially. Moreover, quadratic averaging results are comparable to, if not better than, the performance of the computer alone.

The results of quadratic averaging depend on the number of "training" cases available for the estimation of the bivariate-binormal model parameters and may also depend on the true values of the model parameters. In this study, the number of training cases was on the order of 100. This may be considered as typical in observer studies. Therefore, our results are encouraging because they indicate that the quadratic-averaging method could be practical. However, the parameter values of the bivariate binormal models used in this study were not broad enough for an assessment of the broad effects of quadratic averaging and further studies are needed.

CONCLUSION

Quadratic averaging based on the bivariate binormal model is a theoretically optimal method for combining correlated diagnostic assessments from two sources if the bivariate binormal model is valid and if the model parameters can be estimated reliably. Quadratic averaging may be an effective method for combining radiologists' quantitative diagnostic assessments and a computer's quantitative diagnostic aid to improve on radiologists' *ad hoc* use of the computer aid.

ACKLEDGEMENTS

This work was funded in part by the US Army Medical Research and Materiel Command (DAMD17-00-1-0197), National Institutes of Health (R01-GM57622), and Cancer Research Foundation of America. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of the supporting organizations.

REFERENCES

1. D. J. Getty, R. M. Pickett, C. J. D'Orsi and J. A. Swets, "Enhanced interpretation of diagnostic images," *Invest Radiol* **23**, pp. 240-252, 1988.

2. Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," *Acad Radiol* **6**, pp. 22-33, 1999.

3. H. P. Chan, B. Sahiner, M. A. Helvie, N. Petrick, M. A. Roubidoux, T. E. Wilson, D. D. Adler, C. Paramagul, J. S. Newman and S. Sanjay-Gopal, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study," *Radiology* **212**, pp. 817-827, 1999.

4. Z. Huo, M. L. Giger, C. J. Vyborny, Y. Jiang, R. M. Nishikawa and R. M. Engelmann, "Effectiveness of computer aid for radiologist's classification of mammographic mass lesions," *Radiology* **213 (P)**, pp. 200 (abstract), 1999.

5. C. E. Metz, P.-L. Wang and H. B. Kronman, "A new approach for testing the significance of differences between ROC curves measured from correlated data," in *Information Processing in Medical Imaging,* edited by F. Deconinck, pp. 432-445, Nijhoff, The Hague, 1984.

6. C. E. Metz and J. H. Shen, "Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis," *Med Decis Making* **12**, pp. 60-75, 1992.

7. R. G. Swensson, J. L. King, W. F. Good and D. Gur, "Observer variation and the performance accuracy gained by averaging ratings of abnormality," *Med Phys* **27**, pp. 1920-1933, 2000.

8. D. D. Dorfman, K. S. Berbaum and C. E. Metz, "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method," *Invest Radiol* **27**, pp. 723-731, 1992.

9. P. A. Lachenbruch and M. R. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics* **10**, pp. 1-11, 1968.

10. H. P. Chan, B. Sahiner, R. F. Wagner and N. Petrick, "Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers," *Med Phys* **26**, pp. 2654-2668, 1999.

Computer Classification of Malignant and Benign Microcalcifications in Small-Field Digital Mammograms

YULEI JIANG
ROBERT M. NISHIKAWA
MATTHEW M. MALONEY
MARYELLEN L. GIGER
Kurt Rossmann Laboratories for Radiologic Image Research
Department of Radiology
The University of Chicago, Chicago, Illinois

LUZ L.VENTA
Department of Radiology, Northwestern University, Chicago, Illinois

INTRODUCTION

Mammography is currently the most effect method for breast cancer detection. However, mammography faces challenges to improve its performance in the diagnosis of malignant from benign breast lesions and to reduce the number of biopsy procedures performed on benign lesions. We have previously developed a computer technique to classify clustered microcalcifications in mammograms as malignant or benign. We have shown that this technique can be more accurate than radiologists in differentiating malignant from benign breast lesions (Jiang et al. 1996b). More importantly, we have shown that this technique can be an effective diagnostic aid for radiologists that can lead to improvements in diagnostic performance and biopsy recommendations (Jiang et al. 1999). This computer technique, however, was developed on digitized screen-film mammograms and it has not been extended to full-field digital mammograms (FFDMs). In this study, we apply this computer technique to analyze small-field digital mammograms obtained from a LORAD stereotactic biopsy machine. Our purpose was to evaluate the computer performance in classifying malignant and benign clustered microcalcifications in digital mammograms (Pisano et al. 2000, Nawano et al. 1999).

MATERIALS AND METHODS

LORAD Digital Mammograms

We analyzed mammograms of consecutive biopsies performed in 1997 on a LORAD digital stereotactic biopsy machine at Northwestern University. Of this series, we have obtained biopsy results in 242 cases, of which 61 cases were malignant and 181 cases were benign. These images were obtained during either stereotactic biopsies or needle localization procedures for surgical biopsies. The LORAD machine produces images with two different pixel sizes and almost all images we obtained were 512 × 512 in size. The pixel size of the 512 images was 0.116 mm at the phosphor surface. A few images were 1024 × 1024

in size, and the pixel size of 1k images was 0.058 mm at the phosphor surface. Because of the difference in pixel size, 1k images were not analyzed in this study and one case was eliminated for this reason. For cases of stereotactic biopsy, we analyzed images that were labeled as "scout" and "stereo" views. Other images labeled as "pre-fire," "post-fire," and "post-exam" were not analyzed either because of presence of biopsy instruments or because of absence of microcalcifications post biopsy, but these images were used to help determine the exact biopsy location. Cases that consisted of only scout and stereo views were not analyzed because the exact biopsy location could not be determined reliably. For needle-localization cases, although images were not assigned with different labels as in cases of stereotactic biopsy, we analyzed only those images that would have been labeled as scout views (i.e., without overlying biopsy instruments). Occasionally, a verification image after insertion of a needle or a surgical wire was also analyzed if that particular image depicted the microcalcifications more clearly than the corresponding scout image and if the biopsy instrument did not in any way obscure the microcalcifications. Wires in these images did not affect our computer analysis as long as the wire did not overlap with any microcalcifications because the computer analysis was based entirely on individual microcalcifications.

Computer Classification Technique

We applied the computer technique developed on digitized screen-film mammograms without modification to small-field digital mammograms, except that a linear characteristic curve was used instead of a conventional non-linear Hurter and Driffield (H&D) curve for a screen-film system. This computer technique consisted of an automated feature-extraction stage and a classification stage using an artificial neural network (ANN). For feature extraction, locations of individual microcalcifications were manually identified on a high-quality monitor (Imlogix, St. Louis, MO). Images containing biopsy instruments (needle or wire) were used as reference to determine the exact biopsy location. Eight features were extracted from mammograms: (1) area of a cluster, (2) circularity of a cluster, (3) number of microcalcifications in a cluster, (4) average effective volume of microcalcifications (defined as area times contrast with contrast being converted to units of mm), (5) relative standard deviation in effective volume, (6) relative standard deviation in effective thickness (contrast), (7) average area of microcalcifications, and (8) a shape-irregularity measure that was used to identify linear- or irregular-shape microcalcifications. The calculation of contrast required description of the relationship between, for screen-film mammograms, exposure and film density (i.e., H&D curve) and, for digital images, the relationship between exposure and pixel value. In this study, we assumed a linear relationship for the digital images with a slope of 30-pixel value increment for every 1 mR change in exposure (Roehrig et al. 1993, 1994).

In the classification stage, the computer technique used a conventional feed-forward error-back-propagation ANN to analyze the features and to compute an estimate of the likelihood of malignancy (Jiang et al. 1996b, 1999). As in our previous studies, the leave-one-out method was used to train and evaluate the

ANN (Lachenbruch and Mickey 1968, Fukunaga 1990). The ANN analyzed each image separately; performance of computer classification can therefore be evaluated on a per-image basis. However, to simulate clinical decision-making, computer classification was also evaluated on a per-patient basis by combining classification results from all images from one case into a final assessment. In the per-patient analysis, the final assessment was equal to the highest likelihood-of-malignancy estimate obtained from all images in one case.

Radiologists' Prospective Diagnostic Assessment

All cases in this series included a diagnostic assessment made prospectively by a radiologist at the time of biopsy. These diagnostic assessments were similar to the Breast Imaging Reporting and Data System (BI-RADS) assessment categories: There were five categories (from 1 to 5), with 1 indicating "most likely benign" and 5 indicating "most likely malignant." However, these diagnostic assessments were clearly different from the BI-RADS categories because, according to BI-RADS, category 1 (normal) and category 2 (benign) lesions are not to be recommended for biopsy and therefore would not have been assigned to cases in this consecutive biopsy series. Nevertheless, these diagnostic assessments made by radiologists were invaluable for this study as they allowed us to obtain an ROC curve to quantify radiologists' diagnostic performance and to compare their diagnostic performance with that of computer classification.

Analysis of Classification Performance

We used ROC analysis to evaluate classification performance (Metz 1989). A ROC curve was obtained for radiologists from their diagnostic assessments. Two ROC curves were obtained for computer classification in a per-image and a per-patient analysis. Area under the ROC curve (A_z) and a partial area index, $0.90A_z'$, were used as performance indices (Jiang et al. 1996a). Statistical comparisons were made between two ROC curves of radiologists and of computer classification in the per-patient analysis.

RESULTS

We have analyzed 113 cases thus far. Of these, 34 cases were eliminated for reasons of (1) mass at the biopsy site, (2) images containing only specimen radiographs, (3) all images being 1k in size (different pixel size), or (4) no image showing a needle or a surgical wire (unable to determine the exact biopsy location). Of the remaining 79 cases (176 images), 33 lesions (56 images) were malignant and 46 lesions (120 images) were benign.

The ROC curve obtained from radiologists' prospective diagnostic assessments made at the time of biopsy is shown in figure 1. This radiologists' ROC curve has an A_z of 0.76 ± 0.06 and an $0.90A_z'$ value of 0.21 ± 0.11.

Two ROC curves obtained from computer classification in a per-image and a per-patient analysis are also shown in figure 1. In the per-image analysis, in which each mammogram was analyzed as a separate case, computer classification achieved an A_z of 0.84 ± 0.03 and an $0.90A_z'$ value of 0.25 ± 0.10. In the

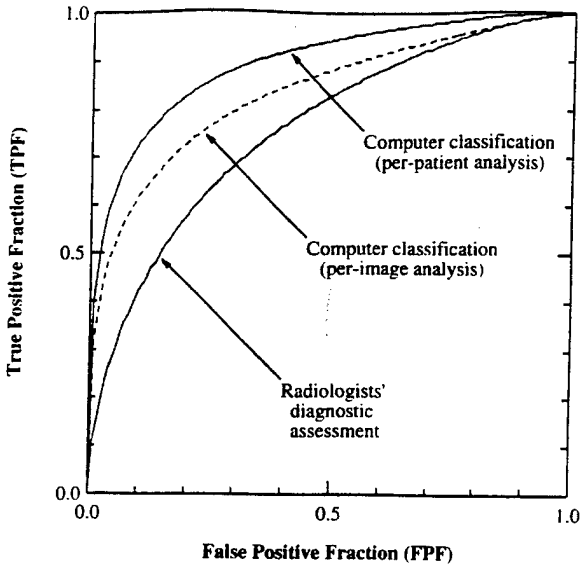

Figure 1. ROC curves comparing radiologists' diagnostic assessment with results of computer classification in a per-image and a per-patient analysis.

per-patient analysis, in which results of all images from one patient were combined into a final assessment by retaining only the highest estimate of likelihood of malignancy, computer classification achieved an A_z of 0.90 ± 0.04 and a $0.90A_z'$ value of 0.44 ± 0.16.

Comparison of two ROC curves between radiologists' performance and computer classification in the per-patient analysis showed that differences in A_z were statistically significant (p = 0.02) but differences in $0.90A_z'$ were not (p = 0.16).

DISCUSSION

The computer performance in classifying malignant and benign clustered microcalcifications obtained from small-field digital mammograms is similar to performance that we have obtained previously from digitized screen-film mammograms. In a previous study of 53 cases of digitized screen-film mammograms, computer classification achieved an A_z of 0.92 (Jiang et al. 1996b). In another study of 104 cases of digitized screen-film mammograms, computer classification achieved an A_z of 0.80 (Jiang et al. 1999). Both studies were designed similar to the present study, and both studies employed the leave-one-out training and evaluation method. In both those studies, the computer performance was found to be significantly better than that of radiologists. These findings indicate that computer performance on small-field digital mammograms is similar to that on digitized screen-film mammograms. However, the results on LORAD digital mammograms are preliminary and need to be updated as we analyze more cases.

We applied the computer technique developed on digitized screen-film mammograms to small-field digital mammograms without modification, except that a linear characteristic curve was used to replace the non-linear H&D curve for screen-film mammograms. In addition to the characteristic curve, our computer technique also uses modulation transfer function (MTF) data (of a screen-film system and a film digitizer) in its calculation of microcalcifications' contrast. Previously, we used MTF data of a Fuji drum scanner (0.1-mm pixel size) and MTF data of a mammography screen-film system (the effect on contrast calculation was dominated by the film-digitizer MTF). In this study, we did not update the MTF data for the LORAD machine because the pixel size of the LORAD machine was comparable to the pixel size of the film digitizer we used previously. However, we will incorporate the correct MTF data (Roehrig et al. 1994) in future studies.

We conclude from this study that our computer technique for classifying clustered microcalcifications as malignant or benign that was developed on digitized screen-film mammograms can be used to analyze small-field digital mammograms. In future studies, we will evaluate this computer classification technique on FFDMs.

ACKNOWLEDGMENTS

The authors thank Dr. Charles E. Metz for his LABROC4 and CLABROC algorithms for fitting and comparing ROC curves. This work was done as part of the International Digital Mammography Development Group (IDMDG). This work was funded in part by NCI through grant CA60187, the U.S. Army through grant DAMD17-00-1-0197, and Cancer Research Foundation of America. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. Robert Nishikawa and Maryellen Giger are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential financial interests that may appear to be affected by the research activities.

REFERENCES

Fukunaga, K. *Introduction To Statistical Pattern Recognition.* Boston: Academic, 1990.

Jiang, Y., C. E. Metz, and R. M. Nishikawa (1996a). "A receiver operating characteristic partial area index for highly sensitive diagnostic tests." *Radiol.* 201: 745–750.

Jiang, Y., R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi (1996b). "Malignant and benign clustered microcalcifications: automated feature analysis and classification." *Radiol.* 198: 671–678.

Jiang, Y., R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi (1999). "Improving breast cancer diagnosis with computer-aided diagnosis." *Acad. Radiol.* 6: 22–33.

Lachenbruch, P. A. and M. R. Mickey (1968). "Estimation of error rates in discriminant analysis." *Technometrics* 10: 1–11.

Metz, C. E. (1989). "Some practical issues of experimental design and data analysis in radiological ROC studies." *Invest. Radiol.* 24: 234–245.

Nawano, S., K. Murakami, N. Moriyama, H. Kobatake, H. Takeo, and K. Shimura (1999 "Computer-aided diagnosis in full digital mammography." *Invest. Radiol.* 34: 310–31

Pisano, E. D., M. J. Yaffe, B. M. Hemminger, R. E. Hendrick, L. T. Niklason, A. D Maidment, C. M. Kimme-Smith, S. A. Feig, E. A. Sickles, and M. P. Braeuning (2000 "Current status of full-field digital mammography." *Acad. Radiol.* 7: 266–280.

Roehrig, H., L. L. Fajardo, and T. Yu (1993). "Digital x-ray cameras for real-time stereo tactic breast needle biopsy." *Proc. SPIE.* 1896: 213–224.

Roehrig, H., L. L. Fajardo, T. Yu, and W. S. Schempp (1994). "Signal, noise and detectiv quantum efficiency in CCD based x-ray imaging systems for use in mammography. *Proc. SPIE.* 2163:320–332.

Computer Classification of Malignant and Benign Microcalcifications in Small-Field Digital Mammograms

YULEI JIANG
ROBERT M. NISHIKAWA
MATTHEW M. MALONEY
MARYELLEN L. GIGER
Kurt Rossmann Laboratories for Radiologic Image Research
Department of Radiology
The University of Chicago, Chicago, Illinois

LUZ L.VENTA
Department of Radiology, Northwestern University, Chicago, Illinois

INTRODUCTION

Mammography is currently the most effect method for breast cancer detection. However, mammography faces challenges to improve its performance in the diagnosis of malignant from benign breast lesions and to reduce the number of biopsy procedures performed on benign lesions. We have previously developed a computer technique to classify clustered microcalcifications in mammograms as malignant or benign. We have shown that this technique can be more accurate than radiologists in differentiating malignant from benign breast lesions (Jiang et al. 1996b). More importantly, we have shown that this technique can be an effective diagnostic aid for radiologists that can lead to improvements in diagnostic performance and biopsy recommendations (Jiang et al. 1999). This computer technique, however, was developed on digitized screen-film mammograms and it has not been extended to full-field digital mammograms (FFDMs). In this study, we apply this computer technique to analyze small-field digital mammograms obtained from a LORAD stereotactic biopsy machine. Our purpose was to evaluate the computer performance in classifying malignant and benign clustered microcalcifications in digital mammograms (Pisano et al. 2000, Nawano et al. 1999).

MATERIALS AND METHODS

LORAD Digital Mammograms

We analyzed mammograms of consecutive biopsies performed in 1997 on a LORAD digital stereotactic biopsy machine at Northwestern University. Of this series, we have obtained biopsy results in 242 cases, of which 61 cases were malignant and 181 cases were benign. These images were obtained during either stereotactic biopsies or needle localization procedures for surgical biopsies. The LORAD machine produces images with two different pixel sizes and almost all images we obtained were 512×512 in size. The pixel size of the 512 images was 0.116 mm at the phosphor surface. A few images were 1024×1024

in size, and the pixel size of 1k images was 0.058 mm at the phosphor surface. Because of the difference in pixel size, 1k images were not analyzed in this study and one case was eliminated for this reason. For cases of stereotactic biopsy, we analyzed images that were labeled as "scout" and "stereo" views. Other images labeled as "pre-fire," "post-fire," and "post-exam" were not analyzed either because of presence of biopsy instruments or because of absence of microcalcifications post biopsy, but these images were used to help determine the exact biopsy location. Cases that consisted of only scout and stereo views were not analyzed because the exact biopsy location could not be determined reliably. For needle-localization cases, although images were not assigned with different labels as in cases of stereotactic biopsy, we analyzed only those images that would have been labeled as scout views (i.e., without overlying biopsy instruments). Occasionally, a verification image after insertion of a needle or a surgical wire was also analyzed if that particular image depicted the microcalcifications more clearly than the corresponding scout image and if the biopsy instrument did not in any way obscure the microcalcifications. Wires in these images did not affect our computer analysis as long as the wire did not overlap with any microcalcifications because the computer analysis was based entirely on individual microcalcifications.

Computer Classification Technique

We applied the computer technique developed on digitized screen-film mammograms without modification to small-field digital mammograms, except that a linear characteristic curve was used instead of a conventional non-linear Hurter and Driffield (H&D) curve for a screen-film system. This computer technique consisted of an automated feature-extraction stage and a classification stage using an artificial neural network (ANN). For feature extraction, locations of individual microcalcifications were manually identified on a high-quality monitor (Imlogix, St. Louis, MO). Images containing biopsy instruments (needle or wire) were used as reference to determine the exact biopsy location. Eight features were extracted from mammograms: (1) area of a cluster, (2) circularity of a cluster, (3) number of microcalcifications in a cluster, (4) average effective volume of microcalcifications (defined as area times contrast with contrast being converted to units of mm), (5) relative standard deviation in effective volume, (6) relative standard deviation in effective thickness (contrast), (7) average area of microcalcifications, and (8) a shape-irregularity measure that was used to identify linear- or irregular-shape microcalcifications. The calculation of contrast required description of the relationship between, for screen-film mammograms, exposure and film density (i.e., H&D curve) and, for digital images, the relationship between exposure and pixel value. In this study, we assumed a linear relationship for the digital images with a slope of 30-pixel value increment for every 1 mR change in exposure (Roehrig et al. 1993, 1994).

In the classification stage, the computer technique used a conventional feed-forward error-back-propagation ANN to analyze the features and to compute an estimate of the likelihood of malignancy (Jiang et al. 1996b, 1999). As in our previous studies, the leave-one-out method was used to train and evaluate the

ANN (Lachenbruch and Mickey 1968, Fukunaga 1990). The ANN analyzed each image separately; performance of computer classification can therefore be evaluated on a per-image basis. However, to simulate clinical decision-making, computer classification was also evaluated on a per-patient basis by combining classification results from all images from one case into a final assessment. In the per-patient analysis, the final assessment was equal to the highest likelihood-of-malignancy estimate obtained from all images in one case.

Radiologists' Prospective Diagnostic Assessment

All cases in this series included a diagnostic assessment made prospectively by a radiologist at the time of biopsy. These diagnostic assessments were similar to the Breast Imaging Reporting and Data System (BI-RADS) assessment categories: There were five categories (from 1 to 5), with 1 indicating "most likely benign" and 5 indicating "most likely malignant." However, these diagnostic assessments were clearly different from the BI-RADS categories because, according to BI-RADS, category 1 (normal) and category 2 (benign) lesions are not to be recommended for biopsy and therefore would not have been assigned to cases in this consecutive biopsy series. Nevertheless, these diagnostic assessments made by radiologists were invaluable for this study as they allowed us to obtain an ROC curve to quantify radiologists' diagnostic performance and to compare their diagnostic performance with that of computer classification.

Analysis of Classification Performance

We used ROC analysis to evaluate classification performance (Metz 1989). A ROC curve was obtained for radiologists from their diagnostic assessments. Two ROC curves were obtained for computer classification in a per-image and a per-patient analysis. Area under the ROC curve (A_z) and a partial area index, $0.90A_z'$, were used as performance indices (Jiang et al. 1996a). Statistical comparisons were made between two ROC curves of radiologists and of computer classification in the per-patient analysis.

RESULTS

We have analyzed 113 cases thus far. Of these, 34 cases were eliminated for reasons of (1) mass at the biopsy site, (2) images containing only specimen radiographs, (3) all images being 1k in size (different pixel size), or (4) no image showing a needle or a surgical wire (unable to determine the exact biopsy location). Of the remaining 79 cases (176 images), 33 lesions (56 images) were malignant and 46 lesions (120 images) were benign.

The ROC curve obtained from radiologists' prospective diagnostic assessments made at the time of biopsy is shown in figure 1. This radiologists' ROC curve has an A_z of 0.76 ± 0.06 and an $0.90A_z'$ value of 0.21 ± 0.11.

Two ROC curves obtained from computer classification in a per-image and a per-patient analysis are also shown in figure 1. In the per-image analysis, in which each mammogram was analyzed as a separate case, computer classification achieved an A_z of 0.84 ± 0.03 and an $0.90A_z'$ value of 0.25 ± 0.10. In the

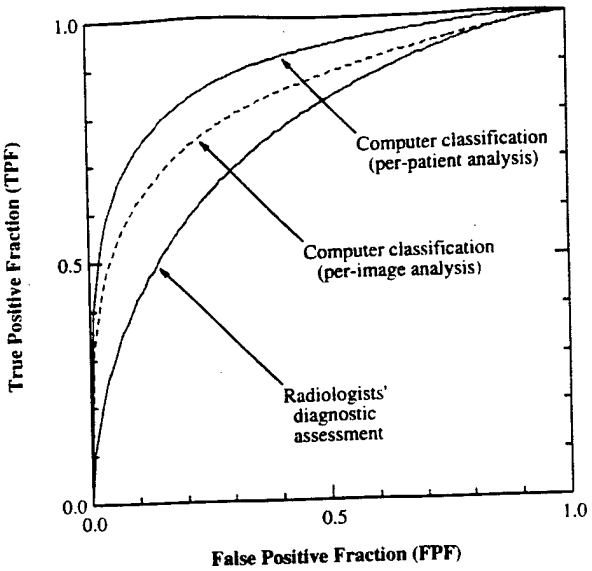

Figure 1. ROC curves comparing radiologists' diagnostic assessment with results of computer classification in a per-image and a per-patient analysis.

per-patient analysis, in which results of all images from one patient were combined into a final assessment by retaining only the highest estimate of likelihood of malignancy, computer classification achieved an A_z of 0.90 ± 0.04 and a $0.90A_z'$ value of 0.44 ± 0.16.

Comparison of two ROC curves between radiologists' performance and computer classification in the per-patient analysis showed that differences in A_z were statistically significant ($p = 0.02$) but differences in $0.90A_z'$ were not ($p = 0.16$).

DISCUSSION

The computer performance in classifying malignant and benign clustered microcalcifications obtained from small-field digital mammograms is similar to performance that we have obtained previously from digitized screen-film mammograms. In a previous study of 53 cases of digitized screen-film mammograms, computer classification achieved an A_z of 0.92 (Jiang et al. 1996b). In another study of 104 cases of digitized screen-film mammograms, computer classification achieved an A_z of 0.80 (Jiang et al. 1999). Both studies were designed similar to the present study, and both studies employed the leave-one-out training and evaluation method. In both those studies, the computer performance was found to be significantly better than that of radiologists. These findings indicate that computer performance on small-field digital mammograms is similar to that on digitized screen-film mammograms. However, the results on LORAD digital mammograms are preliminary and need to be updated as we analyze more cases.

We applied the computer technique developed on digitized screen-film mammograms to small-field digital mammograms without modification, except that a linear characteristic curve was used to replace the non-linear H&D curve for screen-film mammograms. In addition to the characteristic curve, our computer technique also uses modulation transfer function (MTF) data (of a screen-film system and a film digitizer) in its calculation of microcalcifications' contrast. Previously, we used MTF data of a Fuji drum scanner (0.1-mm pixel size) and MTF data of a mammography screen-film system (the effect on contrast calculation was dominated by the film-digitizer MTF). In this study, we did not update the MTF data for the LORAD machine because the pixel size of the LORAD machine was comparable to the pixel size of the film digitizer we used previously. However, we will incorporate the correct MTF data (Roehrig et al. 1994) in future studies.

We conclude from this study that our computer technique for classifying clustered microcalcifications as malignant or benign that was developed on digitized screen-film mammograms can be used to analyze small-field digital mammograms. In future studies, we will evaluate this computer classification technique on FFDMs.

ACKNOWLEDGMENTS

The authors thank Dr. Charles E. Metz for his LABROC4 and CLABROC algorithms for fitting and comparing ROC curves. This work was done as part of the International Digital Mammography Development Group (IDMDG). This work was funded in part by NCI through grant CA60187, the U.S. Army through grant DAMD17-00-1-0197, and Cancer Research Foundation of America. The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. Robert Nishikawa and Maryellen Giger are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential financial interests that may appear to be affected by the research activities.

REFERENCES

Fukunaga, K. *Introduction To Statistical Pattern Recognition.* Boston: Academic, 1990.

Jiang, Y., C. E. Metz, and R. M. Nishikawa (1996a). "A receiver operating characteristic partial area index for highly sensitive diagnostic tests." *Radiol.* 201: 745–750.

Jiang, Y., R. M. Nishikawa, D. E. Wolverton, C. E. Metz, M. L. Giger, R. A. Schmidt, C. J. Vyborny, and K. Doi (1996b). "Malignant and benign clustered microcalcifications: automated feature analysis and classification." *Radiol.* 198: 671–678.

Jiang, Y., R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi (1999). "Improving breast cancer diagnosis with computer-aided diagnosis." *Acad. Radiol.* 6: 22–33.

Lachenbruch, P. A. and M. R. Mickey (1968). "Estimation of error rates in discriminant analysis." *Technometrics* 10: 1–11.

Metz, C. E. (1989). "Some practical issues of experimental design and data analysis in radiological ROC studies." *Invest. Radiol.* 24: 234–245.

Nawano, S., K. Murakami, N. Moriyama, H. Kobatake, H. Takeo, and K. Shimura (1999 "Computer-aided diagnosis in full digital mammography." *Invest. Radiol.* 34: 310–31(

Pisano, E. D., M. J. Yaffe, B. M. Hemminger, R. E. Hendrick, L. T. Niklason, A. L Maidment, C. M. Kimme-Smith, S. A. Feig, E. A. Sickles, and M. P. Braeuning (2000 "Current status of full-field digital mammography." *Acad. Radiol.* 7: 266–280.

Roehrig, H., L. L. Fajardo, and T. Yu (1993). "Digital x-ray cameras for real-time stereo tactic breast needle biopsy." *Proc. SPIE.* 1896: 213–224.

Roehrig, H., L. L. Fajardo, T. Yu, and W. S. Schempp (1994). "Signal, noise and detectiv quantum efficiency in CCD based x-ray imaging systems for use in mammography." *Proc. SPIE.* 2163:320–332.